Developing an Open Source, Reusable Platform for Distributed Collaborative Information Management in the Early Detection Research Network

Andrew F. Hart, Rishi Verma, Chris A. Mattmann, Daniel J. Crichton, Sean Kelly,
Heather Kincaid, Steven Hughes, Paul Ramirez, Cameron Goodale
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109
{hart,rishi.verma,mattmann,crichton}@jpl.nasa.gov

Kristen Anton, Maureen Colbert
Dartmouth Medical School
Lebanon, NH 03766 USA
{kristen.anton,maureen.colbert}@dartmouth.edu

Robert R. Downs
Center for International Earth Science Information Network
Columbia University
61 Route 9W, PO Box 1000, Palisades, NY 10964
rdowns@ciesin.columbia.edu

Christos Patriotis, Sudhir Srivastava
National Cancer Institute
National Institutes of Health
Bethesda, MD 20892, USA
{patriotisc,srivasts}@mail.nih.gov

Abstract

For the past decade, the NASA Jet Propulsion Laboratory, in collaboration with Dartmouth University has served as the center for informatics for the Early Detection Research Network (EDRN). The EDRN is a multi-institution research effort funded by the U.S. National Cancer Institute (NCI) and tasked with identifying and validating biomarkers for the early detection of cancer. As the distributed network has grown, increasingly formal processes have been developed for the acquisition, curation, storage, and dissemination of heterogeneous research information assets, and an informatics infrastructure has emerged. In this paper we discuss the evolution of EDRN informatics, its success as a mechanism for distributed information integration, and the potential sustainability and reuse benefits of emerging efforts to make the platform components themselves open source. We describe our experience transitioning a large closed-source software system to a community-driven, open source project at the Apache Software Foundation, and point to lessons learned that will guide our present efforts to promote the reuse of the EDRN informatics infrastructure by a broader community.

1. Introduction

Cancer ranks among the foremost public health challenges of our time and the quest to better understand the disease continues to consume a tremendous quantity of resources. In 2010 alone, the U.S. National Cancer Institute spent over 16 billion dollars researching the ten most common types of cancer in the United States [1]. That figure has experienced year-over-year increases since 2008 despite the difficult economic conditions of the same period.

Given the scale of the immediate challenges in the field, there is a clear need for efficiency in the development and application of software infrastructures to support ongoing research. This need exists both at the laboratory level as well as at the level of national, collaborative data-sharing networks. Software reuse has the potential to play a significant role in improving efficiency by reducing redundant effort and reinvention, while at the same time providing researchers with increasingly robust tools and services that embrace a distributed, data-intensive model of scientific discovery [12]. Employing an open source model as a vehicle for reuse and sustainability offers opportunities for achieving long-term, organization-independent survivability of software assets, and the potential for their continued growth and evolution via community stewardship.

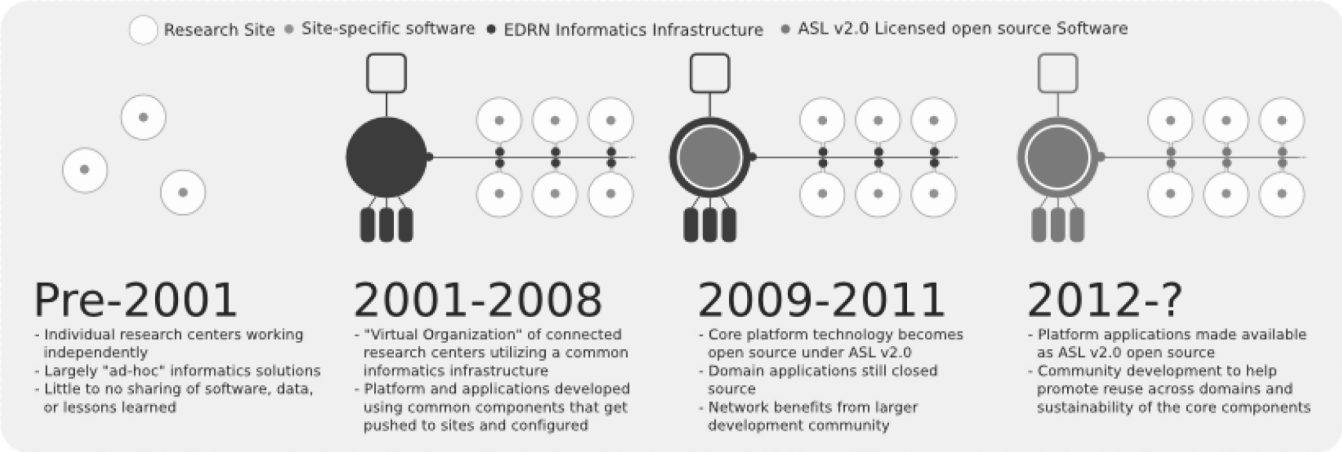

Figure 1. Evolution of EDRN informatics infrastructure over time.

Since 2001, the NASA Jet Propulsion Laboratory (JPL) has served as the center for informatics for the Early Detection Research Network (EDRN) [20], a multi-institution research consortium funded by the U.S. National Cancer Institute. In the intervening years, we have worked closely with the NCI program office and the network's principal investigators to develop a suite of tools and services to support their model of distributed, collaborative scientific discovery. Leveraging an information integration framework that was originally developed at JPL for the purpose of developing reusable data systems for NASA's science missions, we have developed an informatics platform that connects the research activities of the roughly forty institutions across the U.S [6].

During the same period, we successfully transitioned the underlying information integration framework, known as Object Oriented Data Technology (OODT), to the Apache Software Foundation[1] where it holds the distinction of being the first top-level project to originate from within NASA. Transitioning OODT to an open source model was an involved process, which we describe in detail in Section 4.2. The global community of external contributors drives continuous improvement of the software, and helps OODT stand as a clear example of the potential reuse benefits of releasing software under an appropriate open source model.

In part because of our experience with OODT, whose components play integral roles of the EDRN informatics infrastructure, and as a result of interest from similarly structured research organizations in other domains, we have begun the process of transitioning the domain-specific informatics applications for data acquisition, cataloging and archiving, curation, processing, and data dissemination to open source as well. We view this as the first step towards meaningful reuse of the software platform, and see collaboration with other communities as a means to simultaneously improve the platform for EDRN and provide a way to reduce their costs and the time required to develop similar informatics support.

The remainder of this paper is structured as follows. In Section 2 we provide background information on the EDRN and enumerate several related cancer informatics and open-source efforts. In Section 3 we cover the domain-specific applications that comprise the EDRN informatics infrastructure. Section 4 describes our experience with releasing OODT as open source, the lessons learned, and how they apply to EDRN. Section 5 rounds out the paper with a discussion of why we feel the time is right for a transition of the platform to open source, and how stakeholders will benefit.

2. Background

In this section we provide background information about the research activities of the Early Detection Research Network, and briefly discuss related efforts concerning cancer research informatics and open source software development and policy.

2.1. EDRN Research

The Early Detection Research Network is an academic research consortium focused on the identification and validation of biomarkers for the early detection of cancer. It consists of approximately forty principal investigators at various research institutions spread out across the United States, as well as a data management and coordination center at the Fred Hutchinson Cancer Research Center (FHCRC), in Seattle, Washington and an informatics center (IC) at the NASA Jet Propulsion Laboratory (JPL), in

[1] http://apache.org

Pasadena, California.

The EDRN member sites fulfill roles ranging from biomarker development, specimen reference set management, and clinical validation, to study information management, and informatics support. One hallmark of EDRN is the degree of collaboration present between its academic participants. Another is the emphasis placed on informatics in support of research.

2.2. Related Work

The development of EDRN's informatics infrastructure has been an evolutionary process that has progressed from ad-hoc software development at individual sites to more formalized, standardized, and reusable components developed at the IC and deployed across the network. As can be seen from Figure 1, this progression coincides with an increasing focus on the production and use of open-source software. In the following sections, we briefly describe other informatics efforts relevant to cancer research in the context of their reuse efforts. We further outline complementary initiatives being undertaken by the National Aeronautics and Space Administration (NASA) to advance reuse through open source.

2.2.1 caBIG

The Cancer Biomedical Informatics Grid (caBIG) is an NCI-sponsored project seeking a comprehensive bioinformatics grid infrastructure for the cancer research community. Software developed under caBIG funding is made freely available and modifiable, via the terms of their own open source license [21]. Reuse, modification, and even commercialization of software developed by caBIG is possible under these same licensing terms [3]. This open development model has enabled several important collaborations. For example, a partnership between caBIG and Indias Centre for Development of Advanced Computing (C-DAC) is culminating in the formation of the Indo-US Cancer Research Grid (IUCRG) an infrastructure framework leveraging the caBIG open-source codebase [4]. While both caBIG and EDRN seek to simplify access to distributed research information, the EDRN has made a concerted effort to restrict its focus to the specific needs of the early detection community.

2.2.2 BIRN

The Biomedical Informatics Research Network (BIRN) is an effort towards a national data sharing and collaborative capability specifically targeting biomedical research [13]. Funded through the National Institute of General Medicine Sciences (NIGMS) - a unit of the National Institutes of Health (NIH), BIRN offers a range of infrastructural tools and services based in part on the Globus toolkit [9], that includes data sharing utilities as well as workflow pipeline and data cataloging tools. BIRNs software tools and services are made available for both individual researchers and collaborative groups. The infrastructure for collaboration in BIRN is largely decentralized, with the bulk of the software existing at participating sites. This contrasts somewhat with the EDRN's centralized approach, which has permitted the EDRN to retain tighter control over the software deployment and upgrade process.

2.2.3 NASA and Open Source

The U.S. National Aeronautical and Space Agency (NASA) has long held an interest in open source and reusable software [17]. From open source projects such as NASA's World Wind geospatial visualization software [8] to the fact that NASA developed its very own open source license [5], it is evident there is wider appeal to the development and deployment of open source, reusable frameworks within NASA. In fact, the first ever NASA Open Source Summit was held at the Ames Research Center in March, 2011 for the purposes of encouraging agency open source experts to share experiences as well as draft a set of guidelines [2] on the use and development of open source software. The OODT framework, by virtue of its having originated from NASA's Jet Propulsion Laboratory, was examined at the Open Source Summit and found to be consistent with agency recommendations [17]. As we discuss in greater detail in Section 4, EDRN's evolution toward a more reusable and open source platform can be seen as in line with a broader, agency-wide movement at NASA itself.

3. EDRN Informatics

EDRN research is carried out at member institutions that are geographically distant from one another and exhibit significant variation in both human and technological resources. This reality places a premium on the development of a data sharing network that is both capable of supporting the multi-faceted research data and flexible enough to accommodate each of the varying environments it must connect. In this section we describe the major application components of the EDRN informatics infrastructure.

3.1. Overview

The EDRN informatics infrastructure consists of applications and services that provide a layer of connectivity between the activities at each participating site and a network-wide knowledge base. Data management and knowledge transfer between distributed principals forms the heart of the network's informatics aims. As mentioned in Section

2, the applications discussed in this section have evolved as robust replacements for the ad-hoc and often overlapping software development previously required at each site.

The EDRN informatics infrastructure was not developed entirely from scratch. Rather, much of the core data sharing components leverage a pre-existing information integration framework called Object Oriented Data Technology, or OODT, which is described further in Section 4. The flexibility of the OODT framework lies in its component-oriented architecture and this flexibility has allowed for adaptation, customization, and extension of individual components to fit the specific needs of the EDRN's research data environment.

We devote the remainder of this section to describing the major components of the EDRN informatics infrastructure.

3.2. Information Model

One challenge with providing a comprehensive data-sharing environment for the EDRN is accurately representing the multi-faceted nature of EDRN's research data. In the course of identifying and validating cancer biomarker candidates, the EDRN generates a variety of information in the form of protocols, biospecimen collection and processing data, biomarker characteristics, raw and processed datasets, and academic publications describing research outcomes. Taken together, this information provides a rich understanding of the state of research for a particular biomarker.

To facilitate this process, the EDRN maintains an information model and data dictionary that semantically link the major information entities and provide a common, accepted set of terminology for describing the data and its relationships [7]. This open information model is at the core of the EDRN informatics infrastructure, and permeates each of the informatics components, allowing for consistency in knowledge representation between the applications.

3.3. Specimen Data Acquisition

One of the EDRN informatics infrastructure's early successes was the EDRN Resource Network Exchange, or ERNE [14]. ERNE's goal was to remove barriers to obtaining timely and accurate information about biospecimens stored in local repositories at various EDRN sites. Biospecimens in EDRN are collected and stored in various disparate, heterogeneous repositories at individual sites, and information about the contents of each repository is generally maintained using the techniques and technologies available at the host site. While this setup permits researchers at the host site full control over their own repositories, obtaining information about specimens in other repositories usually required direct interaction with someone at that specific site.

ERNE was introduced as a "virtual specimen bank" that enabled a researcher anywhere in the network to quickly ascertain details about specimen resources regardless of their physical location. Using ERNE, a researcher can obtain information about any specimen in the network via a single query. The ERNE application leverages the EDRN common data elements defined in the information model, and transparently issues the query to each connected repository site. ERNE Product Servers installed at the repository sites themselves perform an on-the-fly translation between the ERNE vocabulary and the local data model employed by the repository maintainer at the site. The result from an end-user perspective is a comprehensive result set containing annotated information from all participating sites for specimens matching the given query.

3.4. Data Archiving

EDRN maintains a dataset archive capability that provides a uniform method of storing datasets associated with biomarker research, and a metadata catalog that provides descriptive annotations of the datasets. This dataset repository, the EDRN Catalog and Archive Service, or eCAS, consists of a file-oriented storage system into which raw and processed datasets can be ingested for long-term preservation. The metadata catalog is employed to keep a detailed record of the important facets of each dataset in eCAS, and information about the files in eCAS is made available for searching via a browser-based web interface. The eCAS interface is integrated into the EDRN Public Portal (described in Section 3.6) and supports the ability to explore dataset metadata and download data products either individually or in bulk.

3.5. Data Curation

The reputation of the EDRN as a research network depends strongly upon its ability to consistently produce high-quality information about biomarkers under investigation. To mitigate the challenge of annotating, cataloging, and archiving data arriving from multiple sources, in multiple formats with descriptive metadata annotations of varying degrees of completeness, the informatics center has developed a suite of curation tools to help ensure the consistency, validity, and completeness of the data prior to its public release.

The Biomarker Database (BMDB) is a relational database that stores salient information about each individual biomarker under investigation by the EDRN. This information describes the biomarker itself (e.g.: whether it is genetic, genomic, proteomic, etc.) as well as encodes contextual links to other relevant information (e.g.: organ sites, studies, publications, data sets, etc.). The BMDB

Curation web application [11] is a browser based interface that permits a biomarker curator to annotate the biomarker records as information becomes available and validate the completeness of an individual record.

EDRN data sets archived in the eCAS system described in Section 3.4 also contain significant amounts of contextual metadata critical to understanding how the dataset was generated, what it contains, and how it pertains to the biomarker records in the BMDB. The eCAS Curator web application was developed to facilitate end-to-end dataset information management [10]. The eCAS Curator is both a web application and set of RESTful services for dataset metadata management and file ingestion control. The Curator software is used to maintain complete awareness of what datasets are arriving for ingestion, determine whether required metadata is present or missing, and facilitate the process of archiving a dataset.

3.6. Knowledge Environment

The EDRN informatics infrastructure seeks to provide the EDRN community with a comprehensive "knowledge environment" wherein the extent of the research data products generated within the network can be easily obtained, and relationships between the data can be easily understood. The public face of the EDRN Knowledge Environment (EKE) is the EDRN Public Portal, a web-based interface that provides comprehensive access to the data generated by EDRN in a format that is both context-rich and easy to consume. The EDRN Public Portal takes advantage of the fact that each of the EDRN informatics components subscribes to a common information model and can export data using Resource Description Format (RDF) [15]. In this sense, the Public Portal can be seen as an aggregator, providing up-to-date information regularly collected from across the EDRN enterprise, as well as feature-rich interfaces that promote exploration and utilization of the information.

3.7. Laboratory Support

A more recent addition to the EDRN informatics infrastructure is the Laboratory Catalog and Archive Service (LabCAS). This technology addresses a local laboratory's need to (1) catalog and archive their experimental datasets (2) optionally share experimental datasets across the EDRN in a controlled manner, and (3) leverage customizable data analysis pipelines to gain preliminary insight into these datasets.

One of the most pressing needs LabCAS seeks to address is making available tools to categorize and archive experimental, pre-publication data from individual laboratories. EDRN has built an extensive capability (described in Section 3.4) to catalog and archive peer-reviewed cancer biomarker related datasets. Oftentimes, however, *experimental* datasets generated at the laboratories reside within their own respective data-management systems using dissimilar cataloging techniques. Moreover, sharing and reproducible analysis is made difficult by lack of standardized technologies. The purpose of LabCAS is to provide a standard service for individual laboratories to securely archive their experimental data such that retrieval, analysis, and selective sharing is simplified and standardized.

3.8. A Component Platform

Each of the components described above contributes largely self-contained functionality to the overall informatics platform. While all of these services play a critical role in the context of the EDRN, it should be possible to utilize individual components on a piecemeal basis. This realization is part of the motivation behind emerging efforts to make the informatics components available as open source, which we describe further in the next section.

4. Open Source and Reuse

We hold that the open source model is well suited as a vehicle for promoting software reuse. At the same time, however, "open source" has many meanings, not all of which are equally conducive to developing the kind of vibrant communities needed for an open source project to deliver on this potential [18]. In this section we describe our experience transitioning Object Oriented Data Technology (OODT) to the Apache Software Foundation, and discuss its implications for the EDRN informatics infrastructure.

4.1. OODT

Object Oriented Data Technology [16] is an information integration framework originally developed at JPL out of a growing need for a mission-independent approach to developing largely reusable, large-scale science data systems for NASA missions. Initially funded by the NASA Office of Space Science in 1998, OODT's mandate was the development and implementation of a common, reusable approach to developing data-intensive software systems to accommodate both the growing complexity and volume of the data and the shrinking financial resources devoted to the development of such systems.

A core design goal of the OODT framework is to facilitate greater software reuse. The framework's four main principles 1) division of labor among loosely connected components; 2) well-defined interfaces to guard against internal dependencies on a particular implementation technology; 3) the treatment of metadata (data about data) as

a "first-class" citizen throughout each framework component; and 4) a strict separation of software and data models to allow each to evolve on its own time-line, are all designed to facilitate the ready adaptation of the core platform components to the needs of a particular implementation. This focus on providing modular components and services, and specifically the ability to select and configure individual components to suit has resulted in the framework's broad adoption across a wide variety of disciplines, including planetary and Earth science, space physics, modeling and simulation, pediatric intensive care, and, in the case of the EDRN, cancer research.

4.2. OODT as Open Source

The transition of OODT to the Apache Software Foundation resulted in many lessons learned, several of which we have since codified as part of a broader framework for an agency-wide open source strategy at NASA [17]. The timeline of events can be classified into four major areas.

Process We spent several years working with NASA, JPL, and the California Institute of Technology, and NIH/EDRN program leadership to determine the various dimensions of the release process for open source. This entailed examining OODT for International Traffic and Arms Regulations (ITAR) restrictions (e.g., cryptography, sensitive data and tracking information, etc.), for commercialization opportunities at all levels, and for redistribution restrictions. Once OODT was vetted through this process, we were tasked with determining an appropriate ecosystem for stewardship.

Stewardship We chose the Apache Software Foundation (ASF) as the organization to steward OODT as an open source project. Though initially released through a smaller open source entity called Open Channel, the exposure, active development, release management and triage of projects at Apache were commensurate with our desire to see OODT grow beyond the walls of NASA and into the broader community, and to see it flourish within use cases that our team had never thought of. Furthermore, Apache as a community follows a strict set of processes found to have been effective in the long term maintenance and development of open source software [18].

Community Apache's community consists of over 3000 developers across the world, with projects covering a wide range of areas including WWW/Internet, Databases, User Interface development, libraries, compilers, and browsers, as well as desktop/office software. Apache has an "incubation" process during which new projects are taught the principles of development at the Foundation and important aspects of the Apache open source license. Further, projects are taught to seek sustainability through techniques such as organizational diversity, active addition of new committers and contributors, and frequent software release. Incubating projects are given a chance to learn Apache's meritocratic reporting structure and organization lines, with the goal of developing an independent, sustainable community with clear communication channels and organized leadership via a Project Management Committee (PMC) and a Chair that reports to the Apache board.

Legal During Incubation, Apache projects learn the Foundation's legal policies, including acceptable third party licenses, and permitted upstream and downstream dependencies. Projects are also familiarized with the Foundation's legal structure, provided with free legal open source advice (via the Apache Legal committee), and more.

We heavily leveraged all of the aforementioned resources in OODT's transition to the ASF. During the incubation period, the OODT software was thoroughly filtered through Apache's license conditions, which required us to mitigate multiple conflicts involving third party dependencies with restrictive licenses. In addition, we attracted new, non-NASA contributors to the project (e.g., from Children's Hospital Los Angeles, and from the South African Square Kilometre Array project). The end result of the nearly year-long incubation was an emergence as an independent top-level Apache project with a community that had the essential ingredients to be self-sustaining and operate within the bylaws of the Foundation.

OODT's presence at Apache is of prime benefit to the EDRN community, both in the context of consumption of open source software, and in its production. New EDRN sites that wish to deploy local ERNE and eCAS nodes have the ability to obtain the core OODT software (on which ERNE, eCAS, and other EDRN software depend) from Apache, and easily download, install, and modify the software, for free, with clear and unrestrictive licensing and redistribution policies. On the production side, OODT's successful transition serves a guide for a similar transitioning of EDRN's informatics applications to open source.

4.3. Open Source and EDRN Informatics

The experience of taking OODT from a closed source framework to an open source, community-driven project under the stewardship of the Apache Software Foundation serves as a key point of reference for exploring open source as a vehicle for enhanced reuse of the EDRN informatics software. As a domain-specific implementation of the reference architecture provided by OODT, many of the com-

ponents of the EDRN informatics infrastructure have been designed with the same principles of reuse and adaptability that characterize the framework.

Data acquisition, curation, storage, and dissemination are canonical issues which take on additional importance in the context of a collaborative approach to research where the principal participants are geographically distributed. As described in Section 3.1, the presence of a comprehensive information model that exists independently from the software components that implement the model has facilitated development of purpose-specific applications that nonetheless retain the ability to communicate and fluidly share data with other components in the informatics infrastructure.

We feel that this flexibility is an important feature of the informatics platform. It implies that the tools and services developed in support of the EDRN's challenges may additionally be more generally applicable to similarly structured research communities in other domains. Indeed, our current efforts in this direction are motivated in part by the fact that additional communities have expressed an interest in several of EDRN's informatics components. Providing a direct pathway for obtaining and integrating EDRN informatics applications in other communities via open-source licensing can be mutually beneficial. Other communities will benefit from the ability to adapt and reuse, building off of nearly a decade of lessons learned in the process, and the EDRN, in turn, can benefit from the additional infusion of ideas and innovation brought by increased community involvement.

Finally, adopting an open source model for EDRN informatics offers advantages that extend to the long-term viability of the system in terms of sustainability, reusability, maintainability, and community. An open source platform contributes to sustainability through the elimination of dependencies on vendor-provided modifications, which can result in delays when system improvements are needed [19]. The EDRN informatics infrastructure as it stands today already heavily leverages open source community efforts for its core data sharing components. This implies that the sustainability of the underlying infrastructure is not dependent upon a single organization. By releasing the remainder of the informatics infrastructure as open source, we seek to attain similar assurances for the domain-specific applications that are layered on top.

5. Conclusion

The informatics infrastructure for the Early Detection Research Network (EDRN) has evolved over the past decade from ad-hoc, site-developed software with little or no information sharing or reuse, to a networked platform based on closed-source software developed at the NASA Jet Propulsion Laboratory (JPL), to a mixed platform based on an open source information integration framework from the Apache Software Foundation (ASF) and a robust layer of domain-specific, closed source applications implementing the functionality to support the EDRN's distributed research needs.

As discussed in Section 3, this application layer provides capabilities for data acquisition, curation, metadata cataloging, archiving, processing, and dissemination of the multifaceted research products generated by the network. The need for these capabilities is not unique to the EDRN, however. Similarly structured research organizations interested in facilitating distributed data management among collaborating participants may also encounter the need to support their process with similar capabilities. The EDRN informatics applications are mature, stable implementations that have been operationally proven within the context of a national research network for upwards of five years. The division of labor between the applications and their ability to communicate among each other using common protocols like RDF promises to simplify the process of repurposing them for another domain either individually or as a group.

We believe, as discussed in Section 4.3, that transitioning the domain-specific applications to an open source model will significantly advance the reuse of the software and permit others to leverage the many lessons learned during the development process. As we describe in Section 4.2, the means by which software is made open source can have a direct impact on the nature and quality of the community that develops around the software. Our experience indicates that a healthy, active community is crucial if the expected benefits of reusability and sustainability are to be realized.

At a macro level, what we have developed to meet the EDRN's needs would look similar to what might have developed had the focus had been supporting astrobiology, clinical trials research, or any discipline where a premium is placed on distributed data management with an eye towards public release of high-quality, peer-reviewed information. Our hope is that, as we pursue the open-source release of the EDRN informatics applications, following the precedent set by the successful transition of OODT to the ASF, we can begin to develop a community around the tools and services that promotes both their adoption and reuse in different disciplines and the long-term viability of the applications within the EDRN itself.

Acknowledgements

This effort was supported by the Jet Propulsion Laboratory, managed by the California Institute of Technology under a contract with the National Aeronautics and Space Administration. The authors would like to thank Christos Patriotis, and Sudhir Srivastava, and the NCI leadership as a whole for their collaborative guidance and support.

References

[1] Cancer research funding, http://www.cancer.gov/cancertopics/factsheet/nci/research-funding, 2011.

[2] Nasa open source summit, 2011.

[3] About cabig, http://cabig.cancer.gov/about, 2012.

[4] cabig and india, http://cabig.cancer.gov/action/international/india/, 2012.

[5] Nasa open source agreement v1.3, 2012.

[6] D. Crichton, S. Kelly, C. Mattmann, Q. Xiao, J. S. Hughes, J. Oh, M. Thornquist, D. Johnsey, S. Srivastava, L. Essermann, and W. Bigbee. A distributed information services architecture to support biomarker discovery in early detection of cancer. In *e-Science*, page 44, 2006.

[7] D. J. Crichton, J. S. Hughes, G. J. Downing, H. Kincaid, and S. Srivastava. An interoperable data architecture for data exchange in a biomedical research network. In *CBMS*, pages 65–72, 2001.

[8] D. B. et al. Nasa world wind: Opensource gis for mission operations. *Proc. 2007 IEEE Aerospace Conf.*, pages 1–9, 2011.

[9] I. Foster, C. Kesselman, and S. Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *International Journal of High Performance Computing Applications*, 15(3):200–222, Fall 2001.

[10] A. Hart, C. Mattmann, J. Tran, D. Crichton, J. Hughes, H. Kincaid, S. Kelly, K. Anton, D. Johnsey, and C. Patriotis. Enabling effective curation of cancer biomarker research data. In *Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on*, pages 1 –4, aug. 2009.

[11] A. Hart, J. Tran, D. Crichton, K. Anton, H. Kincaid, S.Kelly, J. Hughes, and C. Mattmann. An extensible biomarker curation approach and software infrastructure for the early detection of cancer. In *Proceedings of the IEEE Intl. Conference on Health Informatics*. IEEE, 2009.

[12] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington, 2009.

[13] D. Keator, J. Grethe, D. Marcus, B. Ozyurt, S. Gadde, S. Murphy, S. Pieper, D. Greve, R. Notestine, H. Bockholt, and P. Papadopoulos. A national human neuroimaging collaboratory enabled by the biomedical informatics research network (birn). *IEEE Trans. Information Technology in Biomedicine*, 12(2):162–172, 2008.

[14] H. Kincaid, S. Kelly, D. J. Crichton, D. Johnsey, M. Winget, and S. Srivastava. A national virtual specimen database for early cancer detection. In *CBMS*, pages 117–123, 2003.

[15] O. Lassila and R. Swick. Resource description framework (rdf) model and syntax specification. Technical report, W3C, 1999.

[16] C. Mattmann, D. J. Crichton, N. Medvidovic, and S. Hughes. A software architecture-based framework for highly distributed and data intensive scientific applications. In *ICSE*, pages 721–730, 2006.

[17] C. A. Mattmann, D. J. Crichton, A. F. Hart, S. C. Kelly, C. E. Goodale, P. Ramirez, and J. S. Hughes. Understanding open source software at nasa. *IEEE ITProfessional*, 14(2):29–35, 2012.

[18] A. Mockus, R. T. Fielding, and J. D. Herbsleb. Two case studies of open source software development: Apache and mozilla. *ACM Trans. Softw. Eng. Methodol.*, 11(3):309–346, July 2002.

[19] M. Schwarz and Y. Takhteyev. Half a century of public software institutions: Open source as a solution to hold-up problem. Working Paper 14946, National Bureau of Economic Research, May 2009.

[20] S. Srivastava and B. Kramer. Early detection cancer research network. *Lab Invest*, 80:11478, 2000.

[21] A. C. von Eschenbach and K. Buetow. Cancer informatics vision: cabig. *Cancer Informatics*, 2:22–24, 2006.